1. Introduction

The coursework aims to prepare a dataset that allows analysing the annual report for NHS Dental statistics in England from 31/3/2018 to 31/3/2019, at the level of Clinical Commissioning Groups (CCGs), which was established in 2012 by the Health and Social Care Act to organise NHS services delivery in England. In addition, analyse the patterns of the number of patients treated in terms of their age, the impact of deprivation, and predict the possible shortage in the number of dentists who aged 55 when they reach State Pension age in 2031, which in turn seems to cause challenging issues to the English CCGs.

There are five datasets that are considered as the raw-data, and they were downloaded and used to construct the final dataset, which contains all the required features to conduct the analysis, those five datasets are respectively:

1. Dentists' data:
   "NHS Dental Statistics for England 2018-19 Annex3_Workforce.csv".
2. Patients' (Jul-Dec 2018):
   "nhs-dent-stat-eng-jul-dec-18-anx3-ps-prac.csv".
3. Patients' (Jan-Jun 2019):
   "nhs-dent-stat-eng-jan-jun-19-anx3-ps-prac.csv".
4. CCG (Mapping of CCG codes to CCG names):
   "nhs-dent-stat-eng-18-19-anx2.xlsx".
5. IMD (index of deprivation):
   "File_13__IoD2019_Clinical_Commissioning_Group*CCG*Summaries.xlsx".

After pre-processing the data (data-cleaning, data-quality related manipulations), we obtained a final .csv file as our final dataset. This dataset is then used for Analysis.

2. Data characterization

The final dataset contains information about patients and dentists in 191 CCGs for different regions in England.

Dataset information:

- Size: 1.2 (MBs)
- Records/Rows: 14516
- Variables: 10

The following list contains the description of each variable:

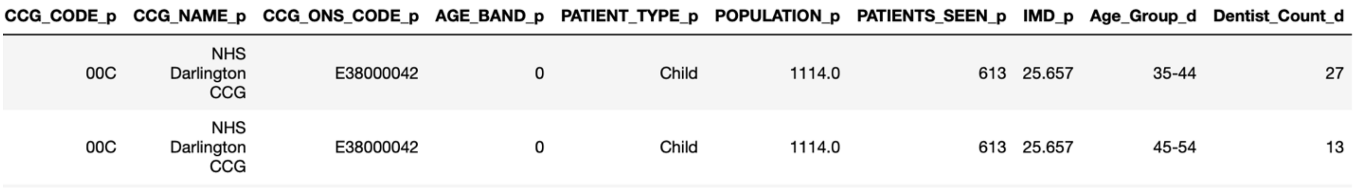
1. CCG_CODE_p: NHS England Clinical Commissioning Group (CCG) Code.
2. CCG_NAME_p: NHS England Clinical Commissioning Group (CCG) Name.
3. CCG_ONS_CODE_p: NHS England Clinical Commissioning Group (CCG) ONS Code.
4. AGE_BAND_p: Age band of patients.
5. PATIENT_TYPE_p: Classification of patient: adult or child.
6. POPULATION_p: The population estimate in the period.
7. PATIENTS_SEEN_p: The number of patients seen in the period.
8. IMD_p: Index of Multiple Deprivation is the average score of population weighted average of the combined scores for the Lower-layer Super Output Areas (LSOAs) in a larger area.
9. Age_Group_d: Age group of dentists.
10. Dentist_Count_d: Count of dentists.

The following table illustrates for each column the data type and statistical data for numerical variables:

Variable name	Data type	Mean	STD	Percentiles				
				Min	25%	50%	75%	Max
POPULATION_p	numerical	15326.379	60114.000	747	2109	3001	4366	962934
PATIENTS_SEEN_p	numerical	94594.685	353591.360	53	13545	21145	31719	5092375
IMD_p	numerical	21.9104	7.974	7.18	16.148	20.826	27.301	52.139
Dentist_Count_d	numerical	43.592	34.456	4	20	34	56	258
CCG_CODE_p	nominal	-	-	-	-	-	-	-
CCG_NAME_p	nominal	-	-	-	-	-	-	-
CCG_ONS_CODE_p	nominal	-	-	-	-	-	-	-
AGE_BAND_p	ordinal	-	-	-	-	-	-	-
PATIENT_TYPE_p	categorical	-	-	-	-	-	-	-
Age_Group_d	ordinal	-	-	-	-	-	-	-

Table 1: The description of data type and missing values and statistical data

A sample of how the data looks like:



CCG_CODE_p	CCG_NAME_p	CCG_ONS_CODE_p	AGE_BAND_p	PATIENT_TYPE_p	POPULATION_p	PATIENTS_SEEN_p	IMD_p	Age_Group_d	Dentist_Count_d
00C	NHS Darlington CCG	E38000042	0	Child	1114.0	613	25.657	35-44	27
00C	NHS Darlington CCG	E38000042	0	Child	1114.0	613	25.657	45-54	13

Table 2: A sample from the final data

3. Data Quality:

The Raw data from the files downloaded were checked for data quality issues. And the following issues were found and fixed.

1. **Redundant data and columns:**
   As our analysis is based on CCG, Patient Age-Band, Dentist Age-Group, Population, Patients count, Dentist count, and IMD-score. The data files were read and all redundant columns were removed in a manner that preserves the consistency of the data.
   a. "*NHS Dental Statistics for England 2018-19 Annex3_Workforce.csv*" – A filter for the Year~'2018-19' and Geography~'4. Clinical Commissioning Group' was applied and only the following Columns were retained in the dataset:
      i. Age_Group
      ii. Dentist_Count
      iii. CCG_CODE
   b. "*nhs-dent-stat-eng-jul-dec-18-anx3-ps-prac.csv"* – A filter was done on the column GEOTYPE~'CCG'. And only the following columns were retained.
      i. CCG_CODE
      ii. CCG_NAME
      iii. CCG_ONS_CODE
      iv. AGE_BAND

   v. PATIENT_TYPE
   vi. POPULATION
   vii. PATIENTS_SEEN

c. *"nhs-dent-stat-eng-jan-jun-19-anx3-ps-prac.csv"* – A filter was done on the column GEOG_TYPE~'CCG'. And the following columns were discarded.

   i. CCG_CODE
   ii. CCG_NAME
   iii. CCG_ONS_CODE
   iv. AGE_BAND
   v. PATIENT_TYPE
   vi. POPULATION
   vii. PATIENTS_SEEN

d. *"File_13_-_IoD2019_Clinical_Commissioning_Group__CCG__Summaries.xlsx"* – As we require only 'IMD – average score' and the mapping columns such as 'Clinical Commissioning Group Code (2019)', we discard the following columns.

   i. *Clinical Commissioning Group Code (2019).*
   ii. *Clinical Commissioning Group Name (2019).*
   iii. *IMD - Average score*

2. **Issues with features/columns name found:**
   - The column name GEOTYPE and GEOG_TYPE, both referring to the Geography type had different identifiers. They were a part of the patients dataset and were later removed after being filtered for 'CCG'.
   - In the dentist dataset, column named 'Org_Code' actually represents the 'CCG_CODE'. So it was renamed to 'CCG_CODE' in order to maintain consistency with other data files.

3. **Nulls:**
   - The raw-data for dentist dataset - 14121 Null values.
     On cleaning and filtering for Geography~4. Clinical Commissioning Group all the Nulls were removed.
   - The raw-data for patients (Jul-Dec 2018) – 1140471 Null values.
     On filtering for GEOTYPE~'CCG' – 124830 Nulls values.
     These nulls were removed after merging the dataset with (Jan-Jun 2019) dataset.
   - The raw-data for patients (Jan-Jun 2019) – 1121280 Null values.
     On filtering for GEOTYPE~'CCG' – 109554 Nulls values.
     These nulls were removed after merging the dataset with (Jul-Dec 2018) dataset.
   - No nulls were found in the IMD dataset.

   Total Null values removed from Raw data – 2275872 Null values.
   After cleaning all the raw-dataset and merging into one final-dataset, another check for Null values were performed and no null values were found.

4. **Special Values:**
   Not many different special values were found. In the raw-data for patients (incl. Jul-Dec 2018 and Jan-Jun 2019) in the CCG columns, there were many 'Unallocated' entries. The 'Unallocated' data had no significant value in our analysis and covered around approx. 6500 records. Therefore was considered a special value and was removed from the final dataset.

5. **Duplicate Values:**
   There were no duplicate records found in the any raw-data.
   After cleaning all the raw-dataset and merging into one final-dataset, another check for duplicate records were performed and no duplicates were found.

4. Data Analysis

The final-data was used to do analysis in order to find answers to the following questions:

a) What patterns are there in the number/age of patients treated?
b) What is the effect of deprivation?
c) A person aged 55 in 2019 will reach the State Pension age in 2031. What types of CCGs face the greatest shortage of dentists in 2031?

a) Patterns in Number/Age of Patients treated:

To get a pattern we divided our analysis on Age of patients and Number of patients into three parts.

1. We check the percentage of patients treated in for Child and Adult. And we check within each band to get some idea about how patients treated are distributed.

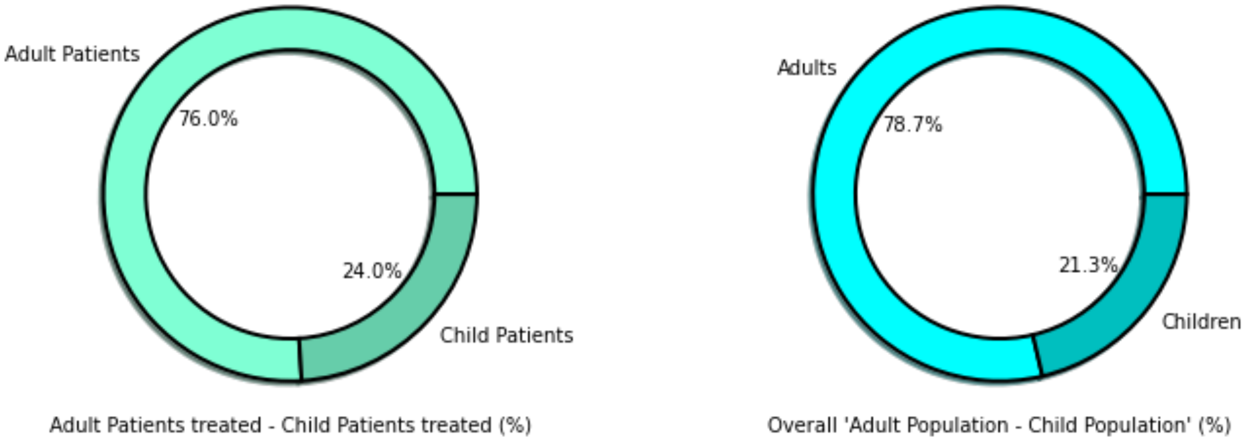


Figure 1: The distribution of Adults and Children patients in the dataset in terms of number treated (left) and overall population (Right).

We can see that there are more Adult patients treated (18+ Age band) as compared to Child patients (0-17 Age bands). This tells us that overall more Adult patients are treated than Child patients. And we can also assume that this could be because of more Adult patients. This could be because of a higher population of Adults in comparison to Children.

2. We analyse Age-Bands of treated patients for children (0 – 17yrs Age) to detect a pattern.

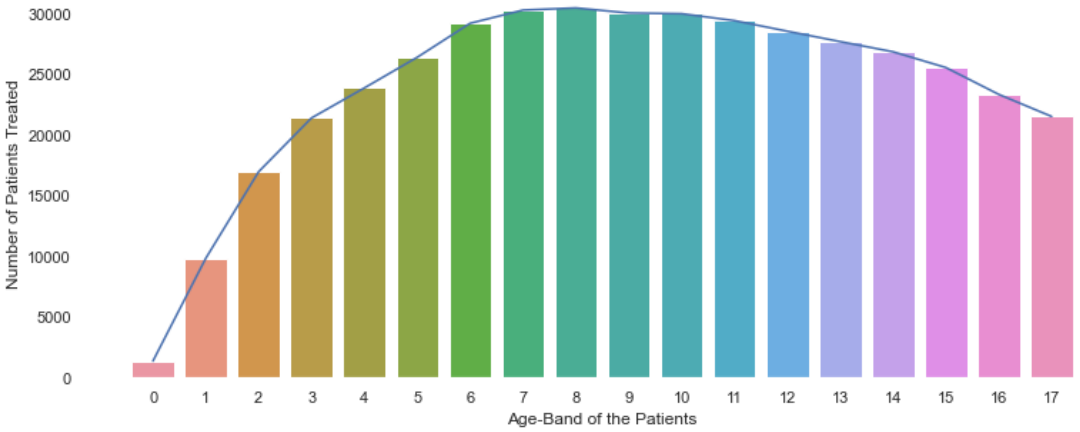


Figure 2: Age-Band of the patients ~ Number of patients seen

For children the plot can be summarized in the following points:

1. An approximate linear increase – The number of patients treated increases with increasing Age-band from Age 0 to 8. So we see that the majority of patients who get treated are those from the age band 8. We can also assume then that there are low number of child patients from the age band 0 to 2 and number of patients increases with the age till the age of 8.
2. A nearly steady plot – A similar case is seen with children from the age between 8 to 10 years of age. These age bands make the majority of treated patients. It can be assumed then that majority child patients come from these age bands.
3. A slight decrease – From Age band 10 we can see a slightly linear decrease in number of treated patients, and it goes all the way till the last age band in children i.e. 17. So we can assume similarly that number of child patients decreases for these age bands.

3. Patients treated per Population comparison for different Age Bands.

To get a meaning out of Patients Band of 18+ which obviously contains ages from 18 and more, we create a a new metric 'Patients treated per Population'. This metric tells us how many patients are treated per population count. We analysed the 'Patients treated per Population' by 'Age-Band' and got the following plot.

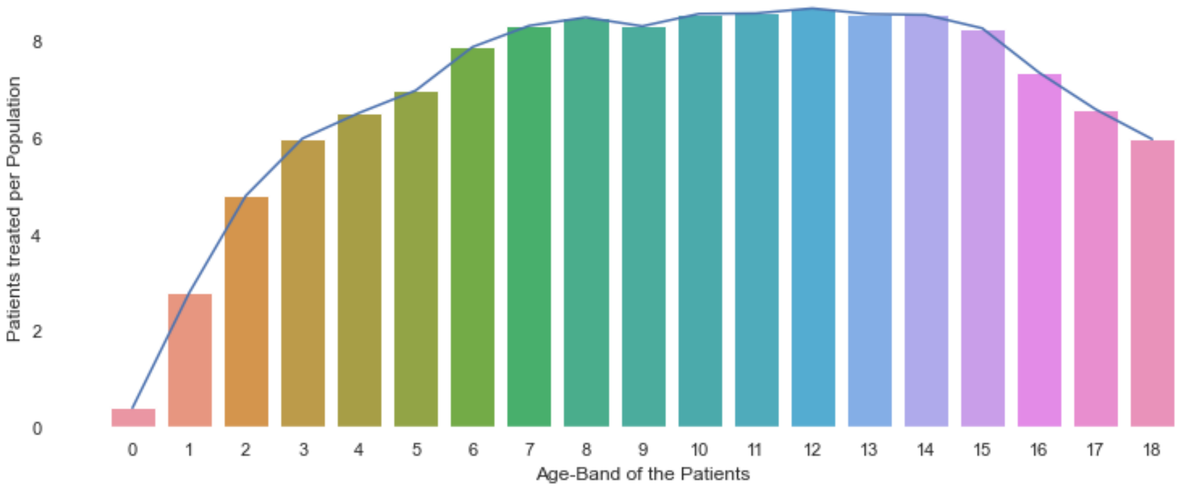


Figure 3: Age-Band of the patients ~ Patients treated per population

Summarizing the pattern into three sections:

1. The pattern of "Patients treated per Population" is very similar to the pattern of "number of child Patients treated – Age Band plot" (plotted above) from Age-Band 0 to 8 years. So we can say that patients treated increases in a what we can refer to as a similar to linear fashion. We can assume that patients are lesser for lower bands than higher bands for these set of bands.
2. From 8 to 16 years the patients treated per population remains steady. So in a population the number of patients treated of these age bands are almost similar or same. This is where it differs from the pattern of "number of child patients treated – Age Band plot".
3. Then we see a steady decline in the bars from age 15 to 18+ bands.

b) Finding an Effect of Deprivation:

On trying to understand the effect of deprivation two features were considered for comparison and analysis.

1. Number Patients seen/treated.
2. IMD (Index of Multiple deprivation) – Average score.

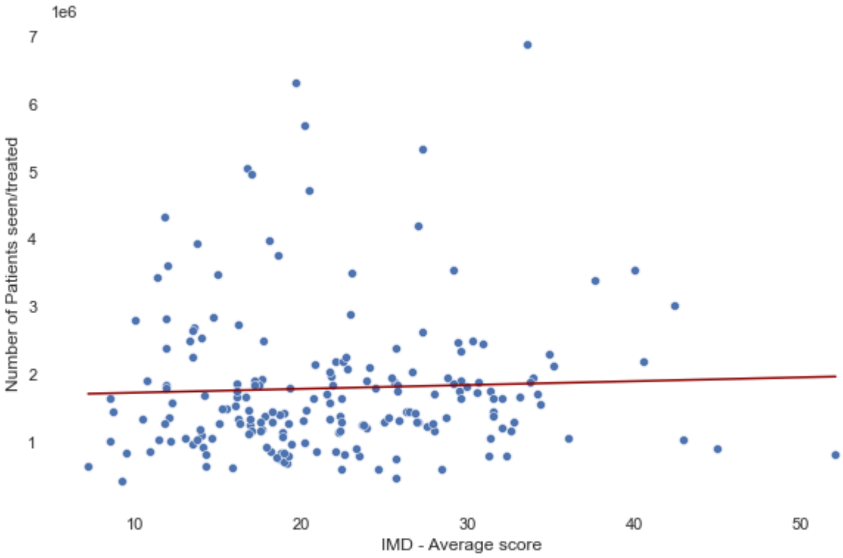


Figure 4: Age-Band of the patients ~ Number of Patients seen/treated.

The scatter plot does not perform very well in establishing a relationship between the two. But we get a gist that IMD does have some effect on the patients seen.

Now we take other two features to take a deeper look into the same:

1. Patients seen/treated per Population.
2. IMD (Index of Multiple deprivation) – Average score.

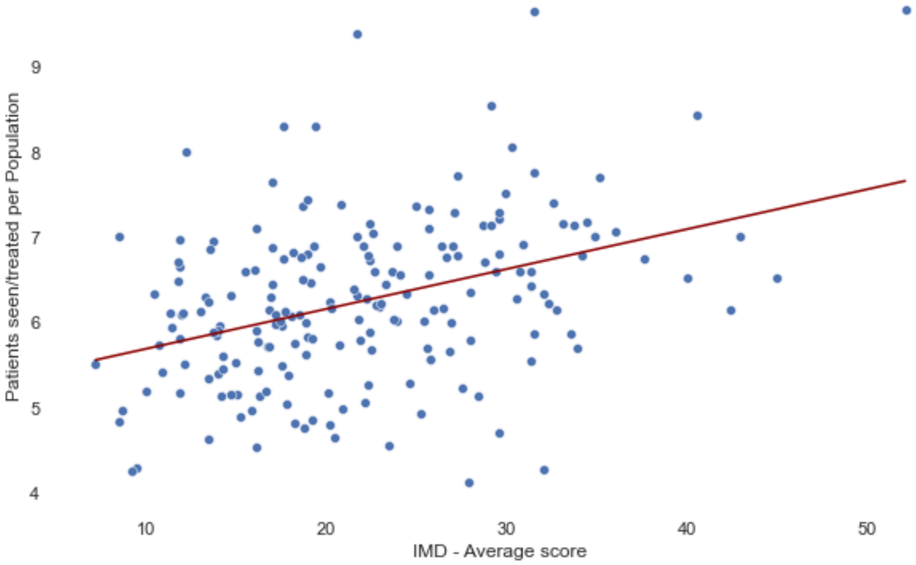


Figure 5: Age-Band of the patients ~ Patients seen/treated per population.

The above scatter plot tells us that there is some correlation between the two features. We can summarise the plot by saying that as Index of deprivation increases from one CCG to other CCG, the Patients treated per population increases as well. In other words, we can assume that this implies to the fact that as Index of deprivation increases so does the number of patients in the area.

We can say that from the effects of deprivation is that an increase in deprivation leads to an increase of patients in an area.

c) A person aged 55 in 2019 will reach the State Pension age in 2031. What types of CCGs face the greatest shortage of dentists in 2031?

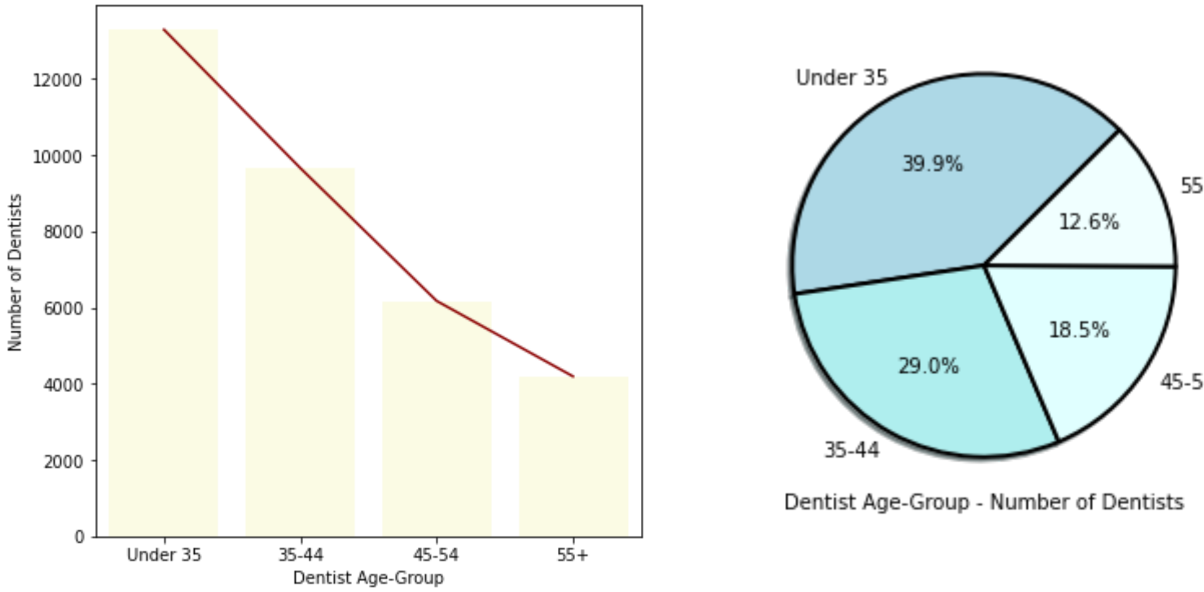


Figure 6: Number of Dentists – Dentist Age-Group.

The graph shows that there is a decrease in number of dentists with the increasing Age-Group. The highest number of dentists was 13286 for the group ‘Under 35’ which is more than three times that of dentists in the Age group ‘55+’.

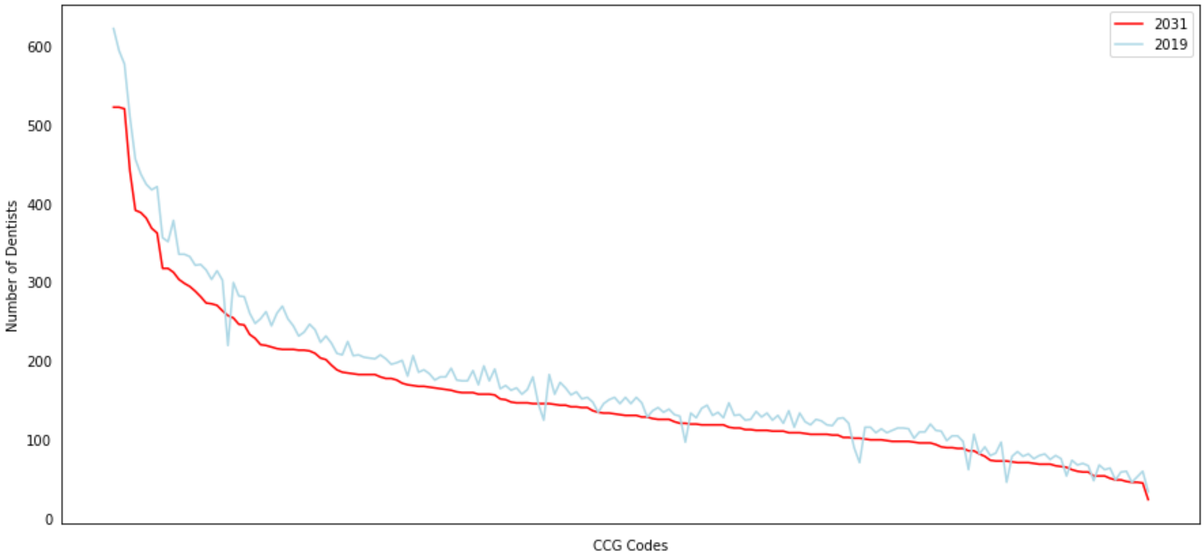


Figure 7: Number of Dentists – CCG Codes (‘2018-19’~’2030-31’)

The graph shows the list of regions with the least number of active dentists in 2019 and 2031. The number of dentists in 2031 is an estimate obtained from the number of dentists under 55 years old in 2019, assuming the number of leavers and joiners remained similar or same.

Dentists in the Age-Group ‘55+’ in the year 2018-19 would reach state pension age and will become in-active. Those from the other Age-Groups will shift one position towards the higher Age-Group.

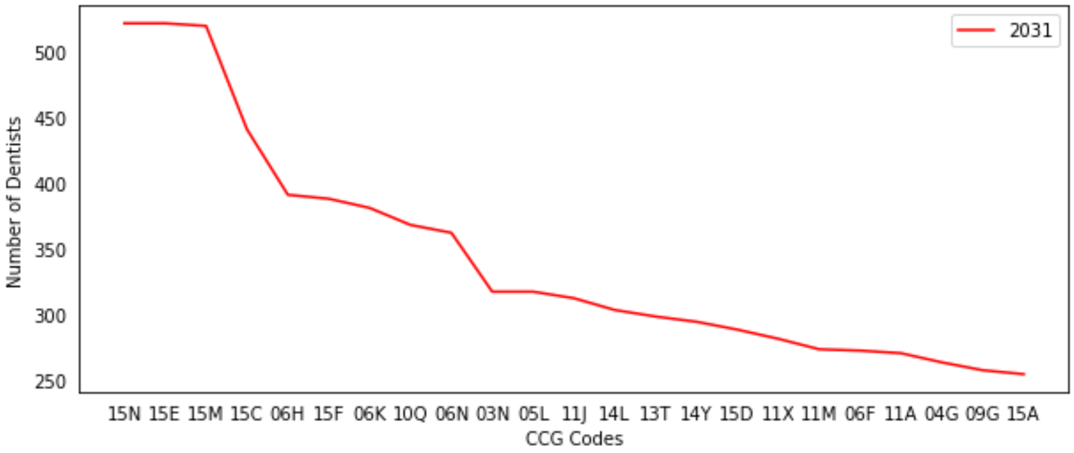


Figure 8: Number of Dentists – CCG Codes ('2030-31')

In 2031, The following regions are the top-5 CCGs to witness a shortage of dentists:

1. NHS Berkshire West CCG (15A)
2. NHS Coastal West Sussex CCG (09G)
3. NHS Nene CCG (04G)
4. NHS West Hampshire CCG (11A)
5. NHS Bedfordshire CCG (06F)

5. Conclusion

The Analysis on the dataset was done and the answers to the queries were found. The main findings of our analysis include the following:

- We saw that highest number of patients treated comes from Age-Band of 8, 9 and 10.
- We saw that the younger children (Age-Band 0 – 2), makes a small proportion of treated patients.
- In a population the number of patients treated of 8-16years age bands are almost similar or same.
- We also saw that it can be assumed that from the effects of deprivation is that an increase in deprivation leads to an increase of patients in an area. As the patients treated and Index of deprivation are slightly linearly related.
- Finally, we calculated and found out the CCGs with the maximum shortage of dentists in 2031. The number of dentists seems to decrease from 2019 – 2031 considering the number of leavers and joiners remained the same or similar.
- Among the CCGs in 2031 "NHS Berkshire West" seems to be the CCG with maximum dentist shortage.